



Behaviors Speak More: Achieving User Authentication Leveraging Facial Activities via mmWave Sensing

Chenxu Jiang
Clemson University
chenxuj@clemson.edu

Sihan Yu
Clemson University
yus@rowan.edu

Jingjing Fu
Clemson University
jfu@g.clemson.edu

Chun-Chih Lin
Clemson University
chunchi@clemson.edu

Huadi Zhu
The University of Texas at Arlington
huadi.zhu@mavs.uta.edu

Xiaolong Ma
Clemson University
xiaolom@clemson.edu

Ming Li
The University of Texas at Arlington
ming.li@uta.edu

Linke Guo
Clemson University
linkeg@clemson.edu

Abstract

Human faces have been widely adopted in many applications and systems requiring a high-security standard. Although face authentication is deemed to be mature nowadays, many existing works have demonstrated not only the privacy leakage of facial information but also the success of spoofing attacks on face biometrics. The critical reason behind this is the failure of liveness detection in biometrics. This work advances most biometric-based user authentication schemes by exploring dynamic biometrics (human facial activities) rather than traditional static biometrics (human faces). Inspired by observations from psychology, we propose the mmFaceID to leverage humans' dynamic facial activities when performing word reading for achieving robust, highly accurate, and effective user authentication via mmWave sensing. By addressing a series of technical challenges of capturing micro-level facial muscle movements using a mmWave sensor, we build a neural network to reconstruct facial activities via estimated expression parameters. Then, unique features can be extracted to enable robust user authentication regardless of relative distances and orientations. We conduct comprehensive experiments on 23 participants to evaluate mmFaceID in terms of distances/orientations, length of word lists, occlusion, and language backgrounds, demonstrating an authentication accuracy of 94.7%. We also extend our evaluation in a real IoT scenario. By speaking real IoT comments, the average authentication accuracy can reach up to 92.28%.

CCS Concepts

• Human-centered computing; • Security and privacy;

Keywords

mmWave, Facial Activity, User Authentication, Biometrics

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Sensys 2024, November 4-7, 2024, Hangzhou, China

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0697-4/24/11

<https://doi.org/10.1145/3666025.3699330>

ACM Reference Format:

Chenxu Jiang, Sihan Yu, Jingjing Fu, Chun-Chih Lin, Huadi Zhu, Xiaolong Ma, Ming Li, and Linke Guo. 2024. *Behaviors Speak More: Achieving User Authentication Leveraging Facial Activities via mmWave Sensing*. In *Proceedings of 22nd ACM Conference on Embedded Networked Sensor Systems (Sensys 2024)*. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3666025.3699330>

1 Introduction

Biometrics-based user authentication has been widely adopted in many sensitive applications and systems, such as online banking [47], healthcare [42], mobile devices [73], voice-controlled smart home [62], and access control in secure facilities [2]. Most of those authentication schemes adopt faces, fingerprints, iris, and voice for identifying authorized users, in which unique biological characteristics can be extracted. However, many recent research works have shown the success of replay attacks to spoof the user's face [18, 59], fingerprint [23, 44], iris [9, 55], and voice [21, 76]. Among those, for example, Apple FaceID leverages the TrueDepth sensor to defend against spoofing attacks using 2D face images, but the 3D-printed face can bypass the face authentication [81]. To further enhance the security level, current practices usually adopt multi-factor authentication by demanding additional biometrics or credentials, e.g., asking users to blink their eyes during authentication in addition to FaceID, which obviously compromises usability. The fundamental reason for this imperfection is that most biometrics are static. Therefore, additional biometrics/credentials are needed to introduce the liveness detection [37, 81], i.e., this user is a real person who follows the instruction of "blink eyes" rather than a 3D-printed face. Existing research works have explored the integration of user authentication and liveness detection (e.g., lip movement [32, 40], blink rate [93], pupil response [94], speech-induce facial vibration[60]). To achieve the authentication, some designs even require sophisticated hardware design [81] or specialized sensors [37, 91]. Even worse, among those using the onboard camera for collecting sensitive facial information as biometrics [32, 37, 91], users' privacy is inevitably compromised.

Motivated by studies [77] in psychology where humans have a better recognition of a face in motion than a static face, we ask **can we design a user authentication scheme that leverages a**

dynamic facial activity as the biometric? Hence, the liveness of the user is automatically verified. Due to its preeminent feature of privacy preservation and capturing minor movements or vibrations, mmWave sensing has shown its advantages in liveness detection [33, 39, 68], for which many commercial smart home devices, such as Aqara Presence Sensor [54], have adopted it for human presence and motion sensing. We assume a smart home scenario as in Fig. 1, the user only needs to perform facial activities (e.g., reading "turn on the light" within seconds) in front of an IoT device equipped with an mmWave sensor. Then, the corresponding smart home application can verify whether she/he is the legitimate user. In particular, each individual's facial activity is driven by a combination of muscle movements in the face area (including lips, chin, jaws, and nose) [41]. Those movements are highly related to genetic factors, cultural backgrounds, and personal experiences [6, 17, 43, 71], leading to individually unique patterns. For example, twins, who share the same genes, may be able to unlock each other's iPhones using Face ID. However, their reactions to a comedy video, such as smirking versus laughing, may vary significantly. Most importantly, while the face contour can be easily replicated, dynamic facial activities are more complex and extremely hard to imitate.

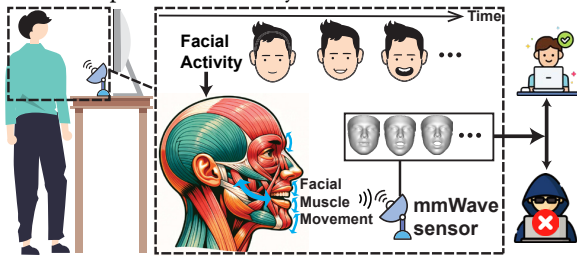


Figure 1: Illustration of mmFaceID System

To capture micro-level muscle movements in facial activities, in this paper, we propose mmFaceID, a mmWave-based user authentication scheme that leverages the newly identified dynamic facial activity biometric. Different from camera or video-based approaches, the proposed mmWave-based method could capture finer-grained facial movements [34], protect privacy and be resilient to low light conditions and occlusions [78], which extends the application scenarios from smart home to healthcare and industrial warehouses where privacy or wearing mask is needed. Our main contributions are as follows,

- Our proposed mmFaceID system validates human facial activity as a behavioral biometric for user authentication without compromising users' privacy.
- We tackle the hardware limitation of the COTS mmWave sensor to implement 2D DoA for capturing fine-grained 3D facial activities.
- Before user authentication, we reconstruct facial activities via estimated facial parameters to make the authentication robust to distances/orientations and achieves high efficiency for new users.
- We design experiments using word reading including real IoT commands to ensure the universality, uniqueness, and permanence of facial activity biometrics. Based on the experiment result from 23 participants, mmFaceID has a 94.7% true positive rate.

2 Motivation and Preliminaries

To design a mmWave-based facial activity biometric authentication scheme, the very first question to ask is *how to choose proper facial*

activities for user authentication? Human facial activities are usually deemed to be unlimited. Even the same person behaves differently on one facial expression under different scenarios. Hence, to ensure the **universality**, **uniqueness**, and **permanence** requirements in biometric authentication, the selected facial activity should 1) be easily performed by any individual; 2) unique to different people; and 3) remain invariant on each individual for a relatively long period (e.g., a few years), respectively. Finally, the proposed biometric must provide a high uncertainty in terms of high entropy to ensure robust authentication.

2.1 Rationale of Facial Activity Biometric

Diverse types of facial activities mainly result from a series of multiple mimetic muscle movements. As shown in Fig. 2, when performing facial activities, the brain triggers the mimetic muscles through a neuromuscular signal that varies [26], resulting in diverse coordination, timing, and strength of the mimetic muscle contractions [13]. Besides, people have different strength levels in their mimetic muscles, which will influence the range and intensity of facial activities [53].

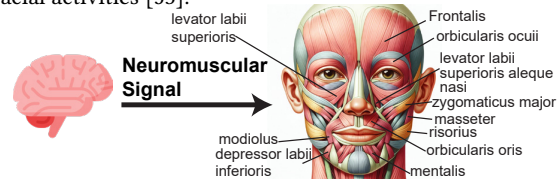


Figure 2: Mimetic Muscles via Neuromuscular Signal

The neuromuscular signals and the corresponding mimetic muscle movements vary significantly among different people mainly from the following three perspectives,

- **Genetic Factor.** The human gene affects the mimetic muscle by the distribution of fiber types, muscle attachment points, neural pathways, neurotransmitters, and bone structure [3, 70]. Taking the distribution of fiber types as an example, slow-twitch fibers contract sustainably, ideal for subtle expressions like a smirk, while fast-twitch fibers generate quick and forceful movements suited for wide grins or expressions of surprise. The distribution of these fiber types is affected by genes and varies between individuals, impacting facial activity intensity and control [57].
- **Cultural Differences.** Different cultures have varying norms and rules about performing facial activities [17]. For instance, some cultures might encourage more expressive displays of emotion, while others might promote more reserved expressions.
- **Personal Experience.** Individual life experiences can also shape how a person performs facial activity. Someone who has experienced certain events may have a unique pattern of expressing emotions on his/her face [6, 43].

In summary, using facial activity as a biometric meets the **uniqueness** requirement, which can be considered for use in an authentication scheme. However, as one of the main sub-categories of facial activities, human facial expressions (e.g., anger, fear, happiness, and surprise) may vary significantly for an individual user even given the same stimuli. Hence, expression-based facial activities are not ideal to be used as biometrics for authentication purposes.

2.2 Motivation

Compared with facial expressions, the facial activities when performing word reading are different [7]. Each word in a language

has a dedicated pronunciation rule that will not frequently change for each individual (i.e., meeting the requirement of **universality** and **permanence**). Besides, each individual has his/her own ways of how to pronounce a word. For example, when producing an "A" sound in "Apple", the person with a stronger orbicularis oris might control his/her lip movements more precisely, affecting the shape and size of his/her mouth opening [61]. Hence, how to capture this kind of **uniqueness** during word speaking becomes the main challenge for user authentication.

2.3 Preliminary Study

To verify our design intuition, we conduct a preliminary study to examine the feasibility of using human facial activity when performing word reading for user authentication.

2.3.1 Data Collection and Analysis. We first use a camera to collect the video when subjects are reading words. Specifically, two subjects are asked to sit in the same position and pronounce some words *Arm, Sharp, Car, Time, Why, Bar* (please refer Sec.6.3 for word selection policy). The distance/angle between the subject and the camera is 30cm/0°.

We adopt 3D Morphable Model (3DMM) [24], a commonly used face model, to analyze the facial activities from the collected video and obtain the expression parameters. The 3DMM model can be described as follows,

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \quad (1)$$

where S is the 3D face mesh, \bar{S} is the mean 3D shape, α_{id} is the shape parameter of the 3D shape base A_{id} (i.e., face contour, not the main focus of this work), A_{exp} is the expression base, and α_{exp} is the expression parameter. Note that A_{exp} and α_{exp} contains no identity information. Any facial activity can be represented as a combination of 29 preset expressions in 3DMM, and the expression parameter is the weight for preset facial expressions [24].

2.3.2 Experimental Results. We first evaluate 1) whether different human facial activities can be captured when reading different words; and 2) whether the performed facial activity is similar among different recording samples. One of the experimental results showing a mouth-related expression parameter from α_{exp} is presented in Fig. 3. To verify the **permanence** characteristic, subjects are asked to perform the same set of expressions/word-reading once on 3 consecutive days. The value of facial expressions vary significantly across days, and thus cannot be used in authentication. In contrast, the value of reading words is almost identical and meets the **permanence** requirement.

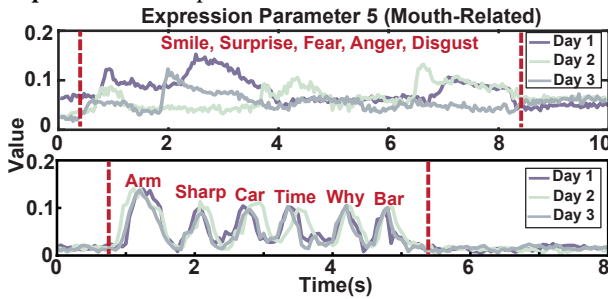


Figure 3: Facial Expression Vs. Word Reading

Taking a step further, we show the **uniqueness** of facial activity biometrics among different subjects. By reading the same words,

the two subjects behave significantly differently in terms of the values in both the expression parameters 5 and 8 as in Fig. 4. Besides, due to different cultural backgrounds, the duration of Subject 2 is longer than Subject 1.

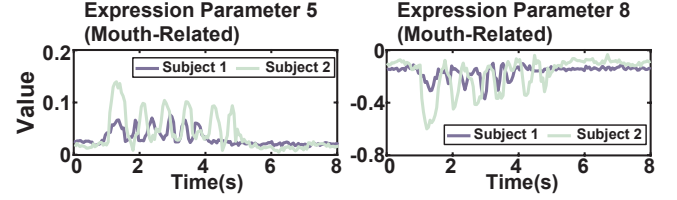


Figure 4: Expression Parameters Comparison

2.3.3 mmWave Sensing-based Validation. We replace the video camera with a mmWave sensor (TI AWR1843BOOST EVM) connected with a data capture card (TI DCA1000EVM) to capture the same facial activities. We use velocity-FFT [36] to compute the velocity of minor facial muscle movements while reading the same words. As shown in Fig. 5, the calculated velocity also demonstrates **uniqueness** and **permanence** of facial activities.

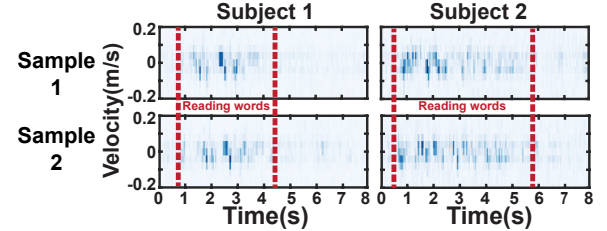


Figure 5: Muscle Movement via a mmWave Sensor

2.4 Entropy Analysis on Facial Activity

Entropy has been widely used to evaluate the security strength of authentication schemes [63]. The classic entropy for a variable x with the distribution $P(x)$ is defined as $H = -\sum_{x \in X} P(x) \log P(x)$. Bringing this to evaluate the biometrics, however, ignores intra-user variability by assuming each user has fixed biometric features and overestimates biometric information. Current research works [1, 63, 85] adopt *relative entropy*, $D(p||q)$, which quantifies the degree of a single user's feature distributions diverge from those of the population [94]. For expression parameter α^i_{exp} , we can extract F features and relative entropy can be defined as,

$$D^i(p||q) = \int_{\mathbf{X}} P(\mathbf{X}) \log_2 (P(\mathbf{X})/Q(\mathbf{X})) d\mathbf{X} \quad (2)$$

where $P(\mathbf{X})$ and $Q(\mathbf{X})$ is feature distributions of an individual and overall population, \mathbf{X} is the feature space of F features. With N_p (individual) and N_q (overall population) samples for feature in F , the feature mean (μ_p, μ_q) and covariance matrix (Σ_p, Σ_q) can be obtained from feature matrix \mathbf{M}_p ($F \times N_p$) and \mathbf{M}_q ($F \times N_q$), respectively. We assume \mathbf{X} follows Gaussian distribution, then $D^i(p||q)$ can be calculated as,

$$D^i(p||q) = (\log_2 \sqrt{e}) \times \left(\alpha + \text{tr} \left((\Sigma_p + \mathbf{T}) \Sigma_q^{-1} - \mathbf{I} \right) \right) \quad (3)$$

where $\alpha = \ln(|\Sigma_q|/|\Sigma_p|)$ and $\mathbf{T} = (\mu_p - \mu_q)(\mu_p - \mu_q)^T$ [63]. As discussed in [1], correlated features are less informative than uncorrelated ones, resulting in a decrease of $D^i(p||q)$. Hence, we extract G ($G \leq F$) mutually independent and important features by Principal

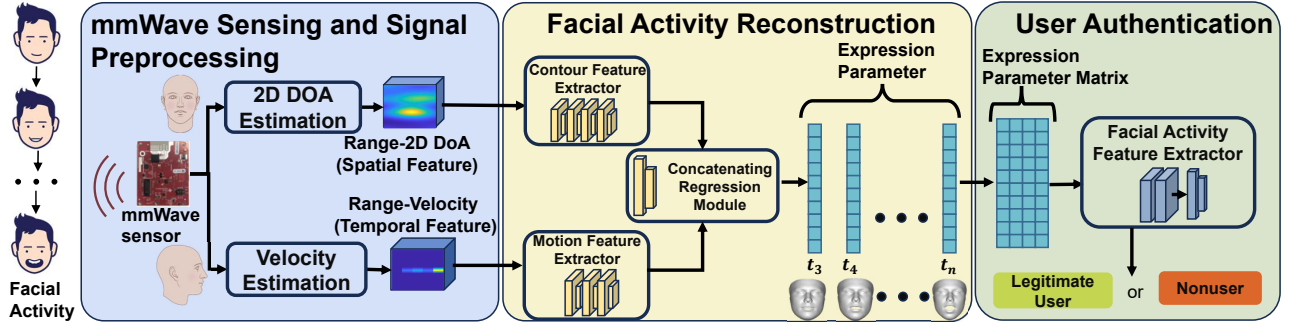


Figure 6: mmFaceID System Overview

Component Analysis (PCA), which can be calculated by Singular Value Decomposition (SVD) as,

$$US_qU^t = \text{svd}(\Sigma_q) \quad (4)$$

where U is orthonormal and S_q is diagonal. The value of S_q indicates the significance of each feature in PCA space. Then, we sort the principal components in descending order of explained variance (σ^2) to determine G as,

$$G = \min \left\{ n \in [1, F] : \sum_{m=1}^n \frac{\sigma_m^2}{\sum_{j=1}^F \sigma_j^2} > th \right\} \quad (5)$$

where $th \in [0, 1]$ is the preset threshold. This equation finds the min n whose cumulative variance ratio is greater than th . Hence, Σ_p is decomposed as $S_p = U^t \Sigma_p U$, and $D^i(p||q)$ can be rewritten in the PCA space as,

$$D^i(p||q) = k \left(\beta + \text{tr} \left(U \left((S_p + S_t) S_q^{-1} - I \right) U^t \right) \right) \quad (6)$$

where $k = \log_2 \sqrt{e}$, $\beta = \ln(|S_q|/|S_p|)$, and $S_t = U^t T U$. Since each expression parameter stands for a unique facial expression and is uncorrelated to each other [25], the relative entropy from N expression parameters can be summed as,

$$D(p||q) = \sum_{i=1}^N D^i(p||q). \quad (7)$$

3 System Overview

By leveraging users' facial activities during word reading, our proposed mmFaceID system can be implemented in many scenarios such as smart home where users' identity can be simultaneously verified when they say some commands/words. The user first needs to train his/her model based on the given commands/words. Then, in order to get authenticated, he/she only needs to speak the pre-trained commands/words within seconds.

3.1 System Framework Design

As shown in Fig. 6, the proposed system mainly consists of the following three modules.

- **mmWave sensing and signal preprocessing.** Traditional 1D DoA can only capture 2D projection of 3D facial activity, resulting unnecessary information lost. We overcome the hardware limitation and innovatively implement 2D DoA on COTS mmWave sensor to acquire the range-azimuth-elevation spectrum of facial activities. After computing 2D DoA and velocity-FFT, the output range-2D DOA and range-velocity are used to extract spatial features and temporal features of facial activities.

- **Facial activity reconstruction.** Existing RF-based face authentication schemes [11, 27] directly use raw signal for authentication,

the accuracy will severely degrade due to the sensitivity of RF signals with respect to distance/orientation. Hence, this module will develop a neural network based on Conv-LSTM to reconstruct facial activity via estimated facial expression parameters. With the range-2D DoA and range-velocity as the input, the output of this module would be a set of expression parameters capturing facial activities.

- **User authentication.** This module aims to identify whether it is the legitimate user from the reconstructed facial activity. The expression parameters estimated from the reconstruction module will be used to extract facial activity features via a CNN-based neural network. Finally, the output of the CNN will be a probability determining whether he/she is a legitimate user or non-user.

3.2 Adversarial Assumptions

To compromise the proposed mmFaceID, the adversary's goal is to impersonate the legitimate user and gain unauthorized access. This work pertains to the discussion of the following commonly seen attacks.

- **Zero-effort attack.** The adversary does not have any side information about the password or the authentication process but will try to get authenticated via random guessing.
- **Replay attack.** The adversary records the video during the authentication process of the legitimate user and replays it to gain access.
- **Shoulder-surfing attack.** The adversary observes the authentication process of the legitimate user. Then, he/she mimics the observed facial activity of the legitimate user.

We also assume the adversary cannot build a 3D dynamic facial activity model from the legitimate user. Other than building a 3D face replica via 3D printing for authentication [81], it is extremely costly to generate a dynamic face model that can adjust facial muscles. Besides, this may need active cooperation from a legitimate user to capture his/her facial movements via specialized devices.

4 mmWave Sensing on Facial Activities

When performing facial activities, the facial muscle moves in a 3D space (spatial feature) and across time (temporal feature). We propose to implement 2D DoA to generate face "imaging" to obtain spatial feature and Velocity-FFT to capture temporal feature.

4.1 Sensing Data Preprocessing

As shown in Fig. 7, the mmWave sensor continuously transmits frequency-modulated continuous wave (FMCW) from the antenna

array to the face. When the subject performs facial activities, the transmitted signal is reflected from different points on the human face and received by the receiver array of the mmWave sensor. The received signal and the transmitted signal are mixed to obtain the intermediate frequency (IF) signal [31]. After collecting the reflected IF signal, we first apply range-FFT to calculate the range of an object, which can be represented as,

$$r = cfT_c/2B \quad (8)$$

where r is the range of the object, and f is the frequency of the IF signal. The value of r is discrete and each possible value is referred to as the range bin [36], where the typical interval between range bins is about 4cm. The result of range-FFT is range-profile depicting the reflected signal from all r . The result of range-FFT (i.e., the range profile) only contains the 1D range information (r) of the object. However, when the facial muscle with different r moves in the 2D facial plane, the dynamic facial activity contains 3D information including range (r), azimuth angle (θ), and elevation angle (ϕ). Hence, we propose to adopt Time-Division Multiplexing (TDM) to achieve 2 Dimension Direction of Arrival (2D DOA), which will depict the shape of the face in a certain moment (spatial feature) and range-velocity to describe the motion of facial muscle movement along with time (temporal feature).

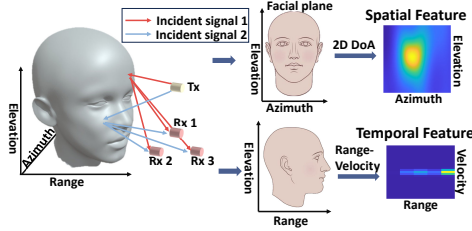


Figure 7: Spatial and Temporal Feature Extraction

4.2 Extracting Spatial Features

4.2.1 Shortcomings of Traditional 1D DoA. Traditional ways (e.g., point clouds [83]) of finding spatial features for depicting the shape of the object usually adopt 1D DOA for each r . The result of 1D DOA is a vector containing signal strength in θ or ϕ . However, the 1D DOA approach assumes only one subject per range bin and cannot sense multiple objects in the same range bin.

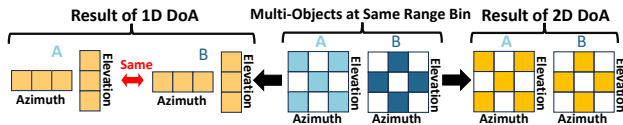


Figure 8: Comparison between 2D DoA and 1D DoA

As illustrated in Fig. 8, A and B are two (multiple) objects in the same range bin with different layouts, the results of using 1D DOA cannot fully picture the Azimuth and Elevation of the original layout, resulting in the same sensing results. On the other hand, the result of 2D DOA approach is a 2D matrix containing Azimuth-Elevation (θ - ϕ). It contains a complete spatial feature of both objects instead of a compressed 1D version using 1D DOA. In other words, θ - ϕ is compressed to θ or ϕ when using 1D DOA, and thus a significant amount of needed information may be lost. Although 1D DOA can output r - ϕ - θ (e.g., point clouds in [83]), it assumes only one object

per range bin. Therefore, all existing designs fail to capture multiple facial muscle movements in the same range bin.

As compared in Tab. 1, current implementations of 2D DOA highly rely on customized devices with complex antenna arrays. Even the TI official document [12] claims to achieve 2D DOA using the COTS mmWave sensor, no existing research works have ever developed any mature schemes due to the lack of detailed code.

| System | Device | DoA | Output |
|---------------------------|--------------|--------|--|
| mmeye [88] | 60GHz Wi-Fi | 2D DoA | θ - ϕ |
| Ren et al. [51] | 2.4GHz Wi-Fi | 2D DoA | θ - ϕ |
| mm3DFace [78] | COTS mmWave | 1D DoA | r - ϕ , r - θ |
| mmMesh [83] | COTS mmWave | 1D DoA | r - ϕ - θ (Point Clouds) |
| m ³ Mesh [82] | COTS mmWave | 1D DoA | r - θ , r - ϕ - θ (Point Clouds) |
| m ³ track [31] | COTS mmWave | 1D DoA | r - ϕ , r - θ , r - ϕ - θ (Point Clouds) |
| mmFaceID | COTS mmWave | 2D DoA | r - ϕ - θ (3D Spectrum) |

Table 1: Comparison among DoA Schemes

4.2.2 Face Scanning via 2D DoA. As shown in the bottom left of Fig. 9, the receiver array of a COTS mmWave sensor is a linear array with four receivers. This physical layout design only allows us to obtain either θ or ϕ , not θ - ϕ , from calculating the angle (of the plane) that is perpendicular to the plane of the receiver antenna array. According to [5], the current 2D layout of the transmitters is the same as a single antenna with more transmission power when transmitting signals, making the 2D DoA infeasible.

• **TDM-induced 2D Virtual Array.** To obtain the 3D matrix containing Range-Azimuth-Elevation using 2D DOA, we propose to create a 2D virtual receiver array by leveraging the time-division multiplexing (TDM) when transmitting mmWave signals [78]. When the TDM is initiated, each Tx takes turns to activate and transmit signals as in the top left of Fig. 9, and all receiver antennas (Rx) are activated during all time slots. By concatenating all the received signals from $0 - 3t$, the virtual receiver design achieves a single Tx with 12 Rx($=4 \text{ Rx} \times 3 \text{ time slots}$). A smaller vertical aperture may raise concerns about lower elevation resolution. In particular, the theoretical elevation angular resolution is described by $\phi_{res} \approx 0.89\lambda/(Nd \cos \phi)$ as in [90]. Here, the antenna spacing is $d = \lambda/2$, antenna number is $N = 2$, and the target angle is $\phi \approx 0$ (the same height with array), so $\phi_{res} \approx 0.89^\circ$. The calculated ϕ_{res} proves the virtual array could achieve the granularity needed for elevation resolution in our scenario.

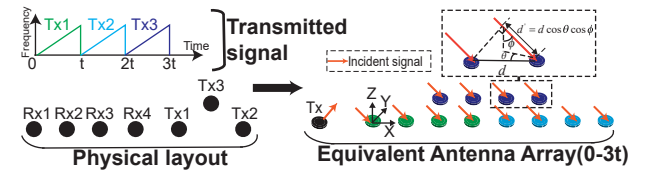


Figure 9: Illustration of 2D DoA

• **2D Array Signal Analysis.** The range between the object and the mmWave sensor (more than 30cm) is much larger than the receiver antenna interval (0.19cm), so the signals arriving at Rx can all be assumed to be parallel. As in the top right of Figure 9, the phase shift ($\Delta\Phi'$) between the signals from two adjacent Rx can be denoted as,

$$\Delta\Phi' = e^{j2\pi \frac{d'}{\lambda}}, \text{ where } d' = d \cos \theta \cos \phi. \quad (9)$$

Note that d' is the traveling distance difference and d is the spacing between two adjacent antennae. Assuming the antenna on the far left of the layout is the reference antenna, for antenna k , the phase shift is,

$$\Delta\Phi_k(\theta, \phi) = e^{j2\pi \frac{d_k(\theta, \phi)}{\lambda}} \quad (10)$$

where d_k is distance difference between the k -th antenna and reference antenna. Therefore, under the noise of \mathbf{N} , the received signal \mathbf{X} by all Rx's from the transmitted signal \mathbf{S} is,

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{N}, \text{ where } \mathbf{A}(\theta, \phi) = [\mathbf{a}(\theta_1, \phi_1), \dots, \mathbf{a}(\theta_j, \phi_j)] \quad (11)$$

Note $\mathbf{a}(\theta, \phi)$ is the steering vector of the antenna array, and $\mathbf{A}(\theta, \phi)$ is the steering matrix for j objects.

• **Spatial Feature Extraction.** We use the two-dimensional minimum variance distortionless response (2D MVDR) (1° angle resolution [50]) to obtain spatial spectrum. In particular, 2D MVDR is a 2D super-resolution DOA estimation algorithm to mitigate the interference and noise from other angles while obtaining distortionless responses to the angle of view [22, 31]. After computing 2D DoA for each range bin, we get the range-2D DoA ($P(r, \theta, \phi)$), which contains the data from all r, θ , and ϕ . In practice, each face lies in 3 range bins (r_f) closest to the preset distance and $[-40^\circ, 40^\circ]$ (θ_f)/ $[-50^\circ, 50^\circ]$ (ϕ_f) in azimuth/elevation. To eliminate unrelated data, we only select the face-related $P(r_f, \theta_f, \phi_f)$, a 3D matrix (tensor), for representing spatial features.

4.3 Extracting Temporal Features

The facial activities causes differences in frequency (Δf) and phase ($\Delta\phi$) between the IF signals at adjacent time slots,

$$\Delta f = 2SvT_c/c, \text{ and } \Delta\phi = 4\pi vT_c/\lambda, \quad (12)$$

where S is the slope of the FMCW signal, v is the object's velocity, and T_c is period of chirps. Since T_c is typically small, Δf is negligible when compared to the frequency of the IF signal [36]. However, $\Delta\phi$ can be detected even with the millimeter-level movement. The movement of facial activity induces the $\Delta\phi$ across the range profile.

To get the temporal features, we first compute Velocity-FFT [36] across the range profile from the consecutive chirps to obtain the range-velocity $RV(r, v)$. Specifically, $RV(r, v)$ is a matrix, in which each element (r_i, v_i) describes the strength of the velocity component v_i of the object (i.e., facial activities) at a range r_i . In practice, facial muscle movement lies in $[-0.2m/s, 0.2m/s]$ (v_f). Similarly, we will only use face-related r_f and v_f to obtain the temporal feature $RV(r_f, v_f)$ for depicting the time-domain-based facial muscle motion.

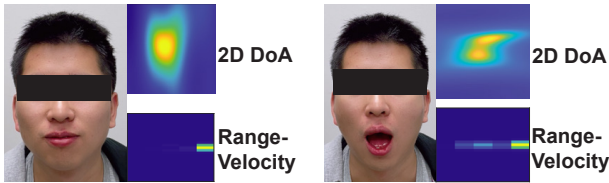


Figure 10: 2D DoA and Range-velocity Comparison

As shown in Fig. 10, after obtaining the face-related spatial feature $P(r_f, \theta_f, \phi_f)$ and temporal feature $RV(r_f, v_f)$ using the above algorithms, the mouth opening/closing can be captured from 2D DoA and range-velocity, demonstrating the facial movement can be uniquely identified from the proposed mmWave sensing approach.

5 Facial Activity Reconstruction

5.1 Reconstruction as a Prerequisite

Most mmWave-based biometric authentication schemes [11, 27] directly use raw mmWave sensing signals for user authentication, which strictly require the user to behave (mostly) in front of the mmWave sensor with a pre-determined distance and orientation. This type of design apparently compromises the usability of authentication schemes. Even worse, it is almost infeasible for those designs to achieve authentication in free space because collecting/training reflected raw mmWave sensing signals at all distances/orientations is infeasible. As shown in Fig. 11, the 2D DoA of the face varies at different distances/orientations, even if those signals should “match” the same person with the same facial activities.

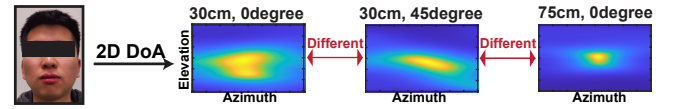


Figure 11: 2D DoA at Different Distances/Orientations

5.2 Estimating Expression Parameters

5.2.1 Deep Learning Framework Overview. Considering the nearly infinite facial activities, we will leverage the 3DMM to estimate the expression parameters for the reconstruction of facial activities. In particular, any facial activity can be represented as a combination of 29 preset expressions in 3DMM [24], in which the expression parameter, α_{exp} , is a vector of weights. Among all 29 high-dimensional parameters in $\alpha_{\text{exp}} \in \mathbb{R}^{29}$, the last 19 parameters have little effect and thus can be ignored. Hence, we only choose the representative first 10 dimensions as $\alpha_{\text{exp}10}$. As shown in Fig. 12, we design a neural network framework to predict $\alpha_{\text{exp}10}$ using the input $P(r_f, \theta_f, \phi_f)$ and $RV(r_f, v_f)$. To train the deep learning model for obtaining the expression parameter, the loss function is designed based on l_1 loss, i.e.,

$$L_E = \text{mean}(\hat{\alpha}_{\text{exp}10} - \alpha_{\text{exp}10}) \quad (13)$$

where $\hat{\alpha}_{\text{exp}10}$ is the ground truth of the facial expression parameter obtained from camera with 3DDFA-V2 [24] and $\alpha_{\text{exp}10}$ is the estimated facial expression parameter.

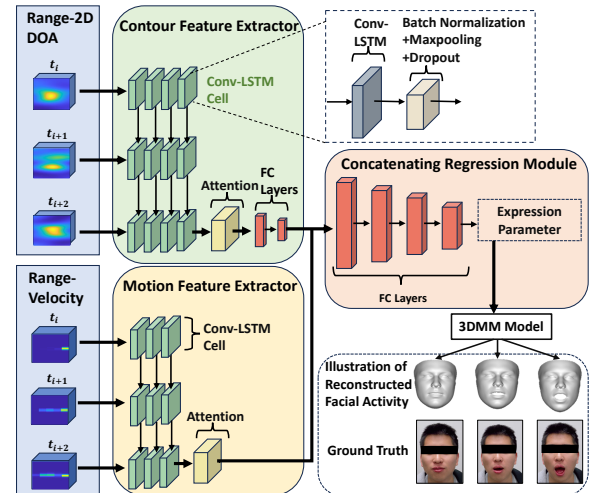


Figure 12: Neural Network Architecture

5.2.2 Feature Extractor. We design the contour feature extractor and motion feature extractor to extract features embedded in the range-2D DoA and range-velocity.

• **Contour Feature Extractor.** This module is designed to learn the face contour embedded in the $P(r_f, \theta_f, \phi_f)$ from facial activities. Usually, the current state of the face contour is highly correlated to the previous state. Meanwhile, the face contour is also regionally correlated with the designated face area. Hence, we use ConvLSTM [56], a combination of CNN and LSTM, as the basis for extracting time-sequential contour features. As shown in Fig. 12, the contour feature extractor consists of three layers, each of which includes four LSTM cells. The feature extraction result will be augmented through the attention module [72]. The output of the attention module goes through two fully connected layers to reduce the feature size. Finally, the feature embedding is stretched using fully connected layers.

• **Motion Feature Extractor.** The motion feature extractor is designed to learn facial muscle movement when performing facial activities. Since the facial muscle movement also changes over time and is regionally correlated, we continue to use Conv-LSTM as the basis of the motion feature extractor. Similarly, this module consists of three layers, each of which includes three Conv-LSTM cells. We can obtain the feature embedding depicting facial motion after going through the attention layer.

5.2.3 Concatenating Regression Module. This module concatenates the feature embedding from the outputs of the two feature extractors for predicting expression parameters $\alpha_{\text{exp}10}$ used in facial activity reconstruction. The concatenating regression module mainly consists of four fully connected layers. Finally, the estimated expression parameters input the 3DMM model to illustrate the reconstructed facial activity. As in Fig. 13, the subject is asked to randomly open/close his mouth while we use a camera and a mmWave sensor to simultaneously record video (as the ground truth) and mmWave signals. The video is inputted into 3DDFA-V2 to obtain expression parameter [24]. The reconstructed activities using facial expression parameters from mmWave sensing are similar to the ground truth.

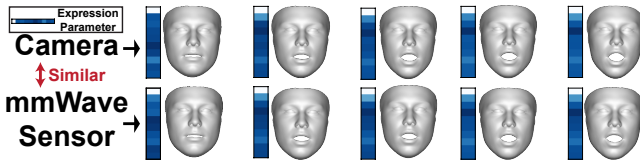


Figure 13: Facial Activity Reconstruction Performance

6 User Authentication

6.1 Building Expression Parameter Matrix

Facial activities when reading a few words in a row normally will last a short period, but the output of expression parameters is based on the consecutive sampling points during the facial activity, e.g., from t_1 to t_3 in Fig. 12. To fully capture unique features used for authentication, it is needed to combine all contour and motion features from the duration of the facial activity [35]. Assuming the user performs the facial activity during a specific period t and the device's sampling rate is f_s , yielding a total of $n(=f_s \times t)$ time points. For each t_i , the range-2D DoA $P(r_f, \theta_f, \phi_f)$ and range-velocity

$RV(r_f, v_f)$ between time point t_i to t_{i+2} will be trained for obtaining the facial expression parameter $\alpha_{\text{exp}10}$, a 10×1 column vector. To capture all facial activities during t , we combine all the $\alpha_{\text{exp}10}$ as the expression parameter matrix A_{exp} , in which the j -th column is $\alpha_{\text{exp}10}$ at t_{j+2} time point.

To validate this, we conduct a similar experiment as in Sec. 2.3.2 where two subjects read the same English words twice in front of a mmWave sensor. The recorded mmWave signals are trained in the above neural network and obtain the expression parameter matrix A_{exp} . Fig. 14 shows that the expression parameter matrices for the same subject are almost identical, whereas those for different subjects behave significantly differently.

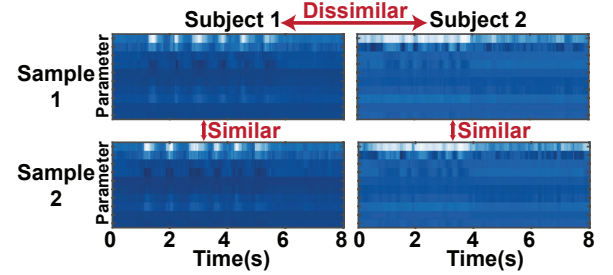


Figure 14: Expression Parameter Matrix, Two Subjects

6.2 Deep Learning Framework Design

We develop a small CNN-based neural network for achieving user authentication as shown in Fig. 15.

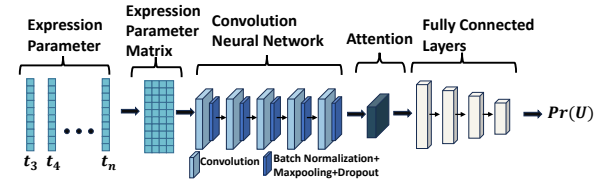


Figure 15: Neural Network Design for Authentication

The expression parameter matrix A_{exp} will be first trained with a CNN (with 5 layers) to extract facial activity features used for authentication. Then, those features will be augmented by the Convolutional Block Attention Module (CBAM) [72], which is used to sequentially infer along two separate dimensions, channel and spatial to obtain attention maps and augmented features. Finally, the augmented features are flattened and inputted into the fully connected layers to estimate the probability of an authenticated user.

6.3 Word Selection Strategy

Our proposed facial activity-based biometric authentication requires users to perform facial movements activated by word reading. To select proper word lists in an application scenario (e.g., "Turn on the Light" in an IoT-enabled smart home), the word list activating more evident facial muscle movements is preferable due to its high entropy. Hence, the selection of proper words used for authentication should jointly consider the requirements/limitations of both sensing and authentication in the dedicated scenario.

Since most word-reading facial activities happen around the mouth area, we investigate the modeling of mouth movements of different phonemes in linguistics. In particular, phonemes are the smallest units of sound in a language that can distinguish one

word from another. Visemes, on the other hand, are the visual representation of a phoneme or group of phonemes when spoken, especially the position and movement of the lips, tongue, and face [16]. Since different phonemes can produce similar visual cues on the user's face, multiple phonemes may be represented by a single viseme. As shown in Fig. 16, two phonemes /b/ and /m/ (last row, right) form a group that maps to a single, near-identical viseme. Therefore, words only with the above two phonemes (e.g., *bat* and *mat*) should not be included in the selected word list. Besides, the selected word should also contain unique phoneme combinations to minimize the similarity.

Instead of focusing on different word lists in diverse application scenarios, we evaluate mmFaceID based on representative visemes, for which we can extend the word list design to more general cases. We choose the most evident mouth movement, viseme *Ah* (first row, left), as the basic viseme, which corresponds to more similar phonemes (other than viseme */MBP/*). Note that Fig. 16 only shows partial of all visemes in [92]. We find the following words meet our rules, "*Buy, Sky, Style, Height, Eye, Shining, Why, High, Guy, Time, Mind, Typing, Arm, Car, Art, Bar, Sharp, Align, Modify, Satisfy*". To jointly consider the similar duration of words during reading, the words "*Arm, Sharp, Car, Time, Why, Bar, Height, Art, Style, Mind, Typing, Guy*" will be used for the word pool for user authentication in this work.







| Viseme | Phoneme | Output | Viseme | Phoneme | Output |
|--------|---------|---|--------|---------------------|---|
| Ah | ɑ, ɔ, a |  | LNTD | l, n, t, d, ʃ, l, r |  |
| Aa | æ |  | GK | g, k, ʃ, q, g |  |
| Eh | e, ɛ |  | MBP | b, m, p |  |

Figure 16: Mouth Movement of Partial Visemes [92]

7 Performance Evaluation

7.1 Evaluation Setup

• **Device Settings.** As shown in Fig. 17, we use a commercial off-the-shelf (COTS) mmWave sensor (Texas Instruments(TI) AWR1843BOOST EVM) connected with a data capture card (TI DCA1000EVM). The mmWave radar has three onboard transmitter antennas and four receiver antennas as in Fig. 9. We configure to transmit mmWave chirp signals on 77-80.984GHz with the signal frame of 468 pulses (156 loops for each Tx) in 50ms. The slope is 40.024MHz/μs. The sampling rate/point is 2,100 kps/200 for each pulse. The range and velocity resolution is 3.93cm and 0.039m/s, respectively. The backend is a PC for reading and processing mmWave signals. We use a camera (Logitech C920 HD PRO) to capture human facial activity and generate facial expression parameters as the ground truth. Note that the camera will only be used for training, NOT in the user authentication process.

• **Data Collection.** We recruited 23 participants (14M/9F) for an IRB-approved experiment (Clemson University Record IRB2023-0688). The participants come from Iran, mainland China, Senegal, US, India, Taiwan, and Bangladesh. The diverse cultural background of the participants helps eliminate the effect of accents when reading words. Participants' ages range from 23 to 38 years old. All data has been anonymized for privacy consideration. During the data

collection, each participant faces both the camera and mmWave sensor while reading words from the developed app as in Fig. 17.

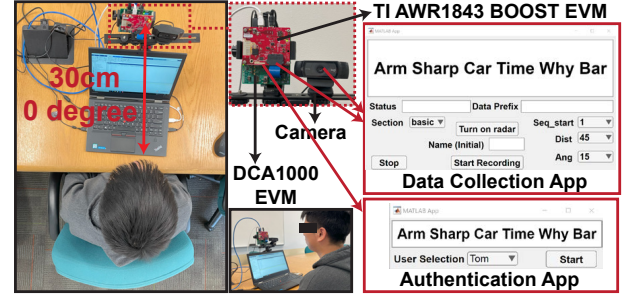


Figure 17: Experiment Settings

All participants will keep their faces still during the data collection. The distance/orientation of the mmWave sensor to the participant's face will vary. Recording time for List1/List2 is set as 8s to ensure every participant can finish. Each participant repeats 10 times for each list at each experiment setting to minimize the bias caused by reading speed with a total duration of 20 min. The file size of each participant's raw data is about 25GB. For each participant, we randomly select 80% of mmWave/camera data to fine-tune a pre-trained reconstruction model from others and the output expression parameter from the trained reconstruction model is employed to train a new authentication model. After training, each user has his/her reconstruction/authentication model. During testing, the remaining 20% of mmWave data from an unknown user is first inputted into the reconstruction model to generate an expression parameter matrix. Then, the generated expression matrix is inputted to the authentication model of the legitimate user for authentication. We implement both neural networks in Pytorch with a dropout rate of 0.5 and a learning rate of 0.0001 using AdamW optimizer on Nvidia RTX4090. We also run the trained neural network on PC and a mobile platform (Google Pixel 7 Pro). As in Tab. 2, our design is efficient in both training (on new users) and inference on PC/mobile platforms during user authentication.

| | Batch Size | Training Time | Inference Time(PC) | Inference Time(Mobile) |
|----------------|------------|--------------------|--------------------|------------------------|
| Reconstruction | 64 | 12.6min (New User) | 279ms | 812ms |
| Authentication | 16 | 3.7min (New User) | 43ms | 137ms |

Table 2: Parameter Setting & Efficiency Analysis

7.2 Performance of Facial Activity Reconstruction

To evaluate the reconstruction, we use the Mean Absolute Error (MAE) to measure the difference between the reconstructed expression parameters and the ground truth. In particular, $n = 10$ is the cardinality of the expression parameter vector. Given the ground truth range (0-6), we find the max MAE is 2.8 obtained from random guessing using the uniform distribution since true distribution is unknown. Note that Max MAE only reflects the worst case and does not affect neural network outputs.

7.2.1 Overall Performance. We first evaluate the reconstruction performance. The distance and angle between the participant and the device is 30cm and 0°. The participant is asked to read the following two lists within 8s.

- **List 1:** *Arm, Sharp, Car, Time, Why, Bar,*
- **List 2:** *Height, Art, Style, Mind, Typing, Guy.*

The average means and mean absolute deviations (MAD) of the reading time are 5.26s (list1)/5.53s (list2) and 0.72s (list1)/0.67s (list2). We observe mouth movements are not directly impacted by the volume. As an example, the reconstructed facial activity when reading “Arm” is shown in Fig. 18. The reconstructed activities successfully show outstanding performance in dealing with subtle facial changes, e.g., gradually enlarging the mouth.

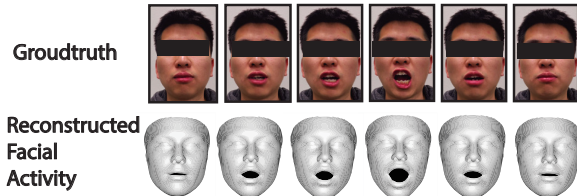


Figure 18: Reconstructed facial activity, “Arm”

In addition to the above empirical result, we use MAE to quantify the reconstruction performance. As in Fig. 19, the MAE of each user when reading both lists is less than 0.25, less than 8.9% of the max MAE. As shown in Fig. 20. The average value for list 1 and list 2 is 0.139, 0.146, respectively. The MAE ranges of the two lists are [0.080, 0.2257] and [0.096, 0.219], respectively.



Figure 19: MAE of Each Participant

Figure 20: Stat.

We take a step further to evaluate the MAE of each reconstructed expression parameter. As shown in Fig. 21, the MAE varies significantly from one parameter to another. Both expression parameters 1 & 2 have a high MAE (but smaller than the max MAE) because those 2 parameters reflect lip pursing/compressing, which causes minimal muscle movement and is hard to be captured by the mmWave sensor. All other larger movements, perfectly captured by mmWave sensing, have low MAE on both lists, demonstrating a high reconstruction accuracy.

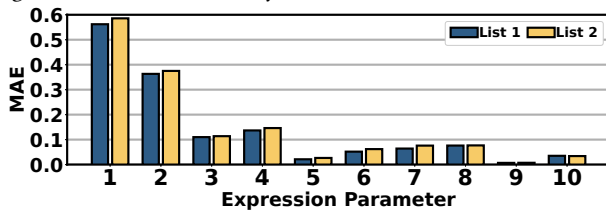


Figure 21: MAE of Each Parameter in α_{exp10}

7.2.2 Impact of the Number of Words. Similar to today’s passwords, increasing the number of words will introduce a higher entropy in the authentication, which usually leads to a lower MAE. By setting the distance as 30cm and the orientation as 0° , we randomly select 1, 3, and 6 words from the word pool. To avoid bias, we choose two sets of words for the set with 3 and 6 words as below,

- **List with 6 words (<6s):** *Arm, Sharp, Car, Time, Why, Bar and Height, Art, Style, Mind, Typing, Guy;*
- **List with 3 words (<3s):** *Arm, Car, Why and Height, Style, Typing;*

- **List with 1 word (<2s):** *Sharp, Art, and Typing.*

As in Fig. 22 and Fig. 23, all the lists have low MAE, indicating the reconstruction can correctly reflect the ground truth regardless of the length of facial activities. Meanwhile, the MAE slightly decreases from 1 word to 6 words, which validates our intuition that more facial activities will help improve the reconstruction performance.

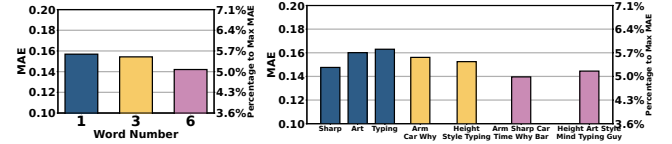
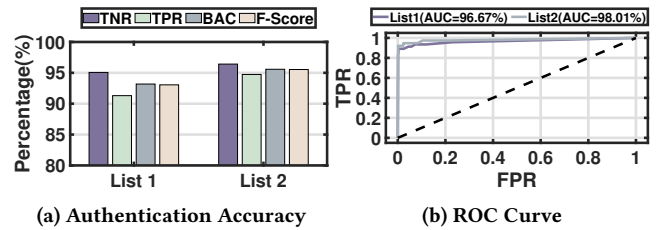


Figure 22: # of words

Figure 23: Different Lists

7.3 Performance of User Authentication

We thoroughly evaluate the authentication performance of the proposed mmFaceID system by considering various types of practical factors. The experimental results from all participants are averaged. We use four metrics to evaluate the performance, i.e., True Positive Rate (TPR), True Negative Rate (TNR), Balanced Accuracy (BAC), and F-Score.

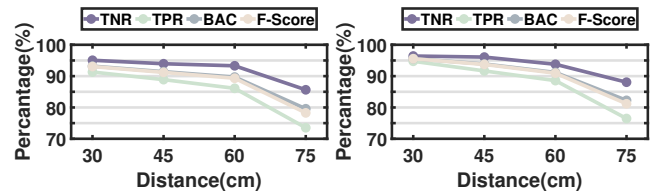


(a) Authentication Accuracy

(b) ROC Curve

Figure 24: Overall Performance

7.3.1 Overall Performance. We first study the overall authentication performance. The distance and orientation between the participant and the device is 30cm and 0° . As in Fig. 24a, the accuracy of the mmFaceID reaches over 90% for all 4 metrics in both word lists. Especially for the average F-Score, List 1 and 2 reach 93.05% and 95.54%, respectively. Note that the result of List 2 is slightly higher than List 1, mainly because words like *Height* and *Style* contain more facial activities to increase the entropy. We also show the Receiver Operating Characteristic (ROC) curve in Fig. 24b, in which Area-Under-Curve (AUC) is 96.67% and 98.01% for List 1 and 2, respectively, showing the developed model correctly classifies all positives and negatives.



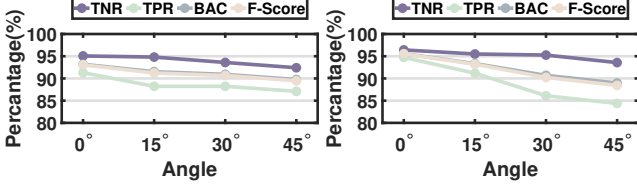
(a) List 1

(b) List 2

Figure 25: The Impact of Distance, 0°

7.3.2 Impact of Distance/Orientation. When increasing the distance from 30cm to 75cm with orientation as 0° , all results in terms of authentication accuracy drop significantly after 60cm mainly due to signal attenuation as in Fig. 25a and Fig. 25b. But, the F-Score still maintains 78.3% and 81.2% for List 1 and List 2, respectively.

By changing the orientation from 0° to 45° at the distance of 30cm, the performance slightly drops as shown in Fig. 26a and Fig. 26b. Compared with the impact of distance, our proposed design is more robust to orientation changes, e.g., the F-score maintains 89.5% and 88.4% at 45° for List 1 and List 2, respectively.



(a) List 1 (b) List 2
Figure 26: The Impact of Orientation, 30cm

7.3.3 Impact of Number of Words. We adopt the same word list setting as in Sec. 7.2.2 and average the results from two sets in both lists with 6 and 3 words. The result of the 1-word list is also averaged from 3 different words. Apparently in Fig. 27, using the 1-word list in mmFaceID, even with our selection rule for ensuring a high entropy, the authentication is not secure because the TPR cannot reach 85%. However, the performance increases steadily when using the 6-word list, which renders a hint of the tradeoff between achieved accuracy level and usability.

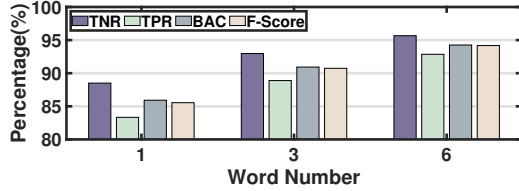
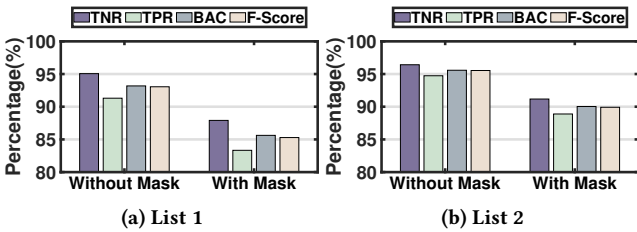


Figure 27: Impact of the Number of Words, 30cm, 0°

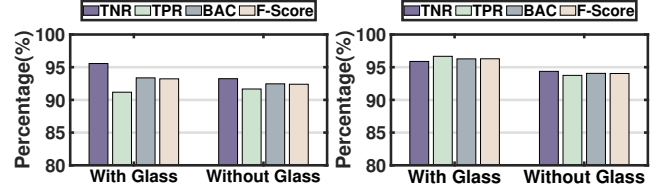
7.3.4 Impact of Mask/Glass. One of the major advantages of using mmWave sensing for authentication is its penetration capability, for which user authentication can be achieved in non-line-of-sight (NLOS) scenarios. As shown in Fig. 28a and Fig. 28b, the performance when wearing a surgical mask (during testing phase) is slightly lower than not wearing it but still reaching 85.3% and 89.9% in F-Score for List 1 and List 2, respectively. Besides the reason for signal attenuation, according to our follow-up survey, participants expressed concerns that the word reading and the corresponding facial activities are restrained when wearing the mask. This result suggests our design may perform better when the mask is not tight to the mouth area.



(a) List 1 (b) List 2
Figure 28: The impact of Mask, 30cm, 0°

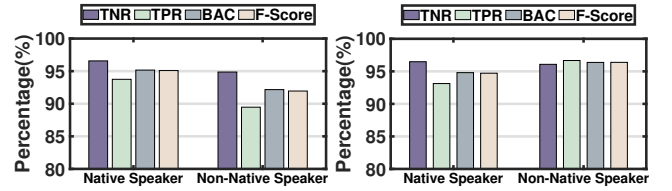
We also ask participants without short or far-sightedness to wear non-prescription glasses when reading word lists. For participants with short-sightedness, they will remove the glasses in this

experiment. As shown in both Fig. 29a and Fig. 29b, no statistical significance can be observed from those word lists, indicating the mmFaceID is robust against glasses compared with wearing a mask. The potential reason is most of the facial activities for word reading do not involve movements in the eye area.



(a) List 1 (b) List 2
Figure 29: The Impact of Glasses, 30cm, 0°

7.3.5 Impact of Language Backgrounds. Our proposed system uses English word-reading for expressing facial activities, for which non-native speakers may have different visemes compared with native speakers on reading the same word. Our experiment population consists of both native and non-native speakers with 5-6 different home languages. Surprisingly, the impact of different language backgrounds is not statistically significant as shown in Fig. 30a and Fig. 30b. As conclusion, our proposed mmFaceID achieves effective and high-accuracy authentication by leveraging unique facial activity signatures from each individual, regardless of how the word is pronounced in different backgrounds.



(a) List 1 (b) List 2
Figure 30: Impact of Language Background, 30cm, 0°

7.4 Comparison With Existing Schemes

As shown in Tab. 3, current face-related authentication systems mainly rely on the camera to extract the face contour, which obviously violates user privacy as in [37, 91, 95]. A recent work [81] leverages a 2D slide rail (20cm×24cm) to scan the face using mmWave radar. Although this design meets all requirements on non-contact, and privacy-preservation, the sophisticated hardware design and the need of fine-grained tuning make the design lack practicality in most of the user authentication scenarios and their working distance is quite shorter than mmFaceID.

7.5 "In-the-Wild" IoT Scenario Evaluation

To further demonstrate mmFaceID can work in the smart home scenario, we invite additional 15 users (9M/6F) to read disruptive IoT commands as shown in Tab. 4 using the same setup in Sec. 7.1.

All the above selected word lists are real IoT command in smart home scenario. As shown in Tab. 4, all BACs are close or over 90%, average BACs is 92.28%, and the TPR can reach to 94.9%, respectively, showing that mmFaceID is secure for real-world IoT commands. In particular, we find that more words in a command

Table 3: Comparison with Current Face-Related Authentication Systems

| System | Medium | Feature | No-Contact | Privacy | Max Distance | TPR(%) | TNR(%) | BAC(%) |
|--------------------|-------------------------|--------------------------------------|------------|---------|--------------|--------|--------|--------|
| LipAuth [32] | Camera | Lip Movement | ✓ | ✗ | 40cm | 94.4 | 99.3 | N/A |
| Li et al. [37] | Camera&Inertial Sensor | Face Contour&Head Pose&Device Motion | ✗ | ✗ | 50cm | 97.7 | 93.9 | N/A |
| Echoprint [91] | Camera&Sound | Face Contour | ✓ | ✗ | 50cm | N/A | N/A | 91.92 |
| Xu et al. [81] | mmWave | Face Contour | ✓ | ✓ | 20cm | 96 | N/A | N/A |
| Jawthenticate [60] | Inertial Motion Sensors | Speech Facial Vibration | ✗ | ✓ | 0cm | 92 | 93 | 92 |
| mmFaceID | mmWave | Facial Muscle Movement | ✓ | ✓ | 75cm | 94.7 | 96.4 | 95.6 |

does not usually guarantee better authentication performance. The reason is that some words in a command contain multiple evident visemes, which are sufficient enough to provide a high level of entropy used for authentication, further showcasing the viseme-based word selection strategy is sound.

| Command | TNR(%) | TPR(%) | BAC(%) |
|----------------------------|--------|--------|--------|
| Unlock Kitchen Window | 94.4 | 91.7 | 93.0 |
| Turn On the Light | 92.1 | 86.4 | 89.3 |
| Open the Front Door | 92.8 | 91.5 | 92.1 |
| Delete All the Data | 93.8 | 88.3 | 91.0 |
| Disable the Security Alarm | 97.0 | 94.9 | 96.0 |

Table 4: Performance of real IoT commands

7.6 Facial Activity Entropy Analysis

We evaluate the entropy (in Sec. 2.4) of the proposed facial activity biometric using reconstructed expression parameters in Sec. 7.2.1. We adopt commonly used 27 statistical features for calculation including mean, maximum/minimum and its index, variance, skewness, kurtosis, interquartile range, mean crossing, root mean square, crest factor and slope sign change in time/frequency domain [38, 45, 93].

| Work | Authentication | Entropy(bits) |
|------------------------|-----------------|----------------------------------|
| Wang et al. [66] | Password | 20-23 |
| Wang et al. [67] | PIN | 8.41(4-digit), 13.21(6-digit) |
| Sae-Bae et al.[52] | Keystroke | 3.48-4.62 |
| Takahashi et al. [64] | Fingerprint | 18.6 |
| Inthavisas et al. [29] | Voice | 18-30 |
| Adler et al. [1] | Face | 37.0-55.6 |
| mmFaceID | Facial Activity | 75.4(List 1)/78.7(List 2) |

Table 5: Entropy of various authentication methods

We set the threshold th as 0.95 in our evaluation. Tab. 5 compares the entropy of various authentication methods. Our proposed mmFaceID has the highest relative entropy (> 75.4 bits) compared with classic non-biometrics (e.g., PIN and password) and traditional biometrics (e.g., face, voice, and fingerprint), showcasing facial

activity biometrics bear a high uncertainty to be used for user authentication.

8 Robustness Against Attacks

In corresponding to our threat model, we continue to conduct experiments to demonstrate the resilience of mmFaceID under different attacks. Specifically, we define Defense Success Rate (DSR) to evaluate the probability that an illegitimate authentication request (via different attacks) is successfully detected. We will use the word list introduced in Sec. 7.2.2 for the evaluation. The results are shown in Tab. 6.

| Words | Zero Effort Attack | Shoulder-surfing Attack | Replay Attack |
|----------------------------|--------------------|-------------------------|---------------|
| <i>Sharp</i> | 95.6 | 88.4 | 99.2 |
| <i>Art</i> | 94.8 | 87.6 | 100 |
| <i>Typing</i> | 92.4 | 90.4 | 97.3 |
| <i>Arm/Car/Why</i> | 98.0 | 96.4 | 100 |
| <i>Height/Style/Typing</i> | 97.6 | 91.8 | 99.3 |
| <i>List 1</i> | 100 | 96.8 | 99.1 |
| <i>List 2</i> | 99.2 | 94.4 | 100 |

Table 6: DSR(%) Under Different Attacks, 30cm, 0°

• **Zero-effort attack.** Zero-effort attackers attempt to pass the authentication by randomly performing facial activities. We randomly select 5 participants as victims and 10 participants as attackers. Attackers perform 5 attempts by saying random words for each word list. Please note that TNR in Sec. 7.3.1 is obtained when saying same words, which is different from zero-effort attack. For the 1-word list, the DSR is above 90% for a single word while reaching over 99% for the 6-word list, indicating that mmFaceID is robust to the zero-effort attack.

• **Shoulder-surfing attack.** We select 10 participants to perform the shoulder-surfing attack on the other 5 participants, i.e., mimicking others' visemes via recorded videos. The attacker will try to access mmFaceID repeatedly for 5 times on each word list. The DSR achieves over 87% for 1-word lists and keeps increasing when the length of the word list increases. As stated in Sec. 7.3.3, more words in the list lead to a higher entropy in authentication.

• **Replay attack.** We use a smartphone (iPhone 12 Pro) to replay the recorded video 10 times for each word list from 10 participants. All DSR is above 98% showing that the 2D video cannot pass our

authentication system. The main reason is our mmWave sensing-based design requires the input to be 3D face, for which the 2D video does not contain 3D facial contour/motion used in the authentication scheme.

The above analysis proves mmFaceID is robust against attacks and a secure authentication method because it leverages diverse facial activities that are hard to fake.

9 User Study

We ask each participant to answer 4 follow-up questions (from 1 strongly disagree to 5 strongly agree) on how they feel about mmFaceID system after they finish the experiment. As shown in Tab. 7, around 91.3% of the participants do not express any concerns when using the mmWave sensor (Q1). As one of the main features of using mmWave sensing other than video, 78.2% of participants agree that their facial information will not be leaked with mmFaceID (Q2). Many participants agree that mmFaceID has potential to replace current biometric-based authentication with positive comments such as "mmFaceID is interesting and good" and "I would like to use it on my laptop".

| Question | Average Score | Percentage(Agree& Strongly Agree) |
|---|---------------|-----------------------------------|
| No discomfort when using mmWave sensor | 4.6 | 91.3% |
| Compared with camera, mmWave sensor can better protect my privacy | 4.3 | 78.2% |
| mmFaceID is easy to use | 4.4 | 86.8% |
| I would like to try mmFaceID in the future | 4.2 | 82.6% |

Table 7: Survey Questions Responses

10 Related Work

• **Behavior-based User Authentication.** Many behavioral biometrics have been used for user authentication in literature, e.g., gaits, touching, voice, and gestures. Gait analysis has been widely used in user authentication schemes that leverage the individual's unique walking behaviors via RF signal [14, 30, 51], camera [35, 58] or wearable sensor [4, 20]. When interacting with devices with touchable screens or keyboards, people behave differently, e.g., touching gesture [46], tapping interval [28], touch-induced vibration [84], and keystroke dynamics [8, 49]. Voice-based authentication employs features when speaking, such as vocal vibrations[34], speech-induced facial vibration[60] and lip movement[87]. Besides, gesture-based authentication exploits unique features when performing gestures via depth camera [69], RF signals [89], and wearable sensors [86]. However, the above studies either lack privacy preservation (e.g., using cameras for collecting behaviors) or introduce additional noises to the surroundings. In contrast, mmFaceID protects privacy and doesn't introduce noise.

• **Human Face Reconstruction.** Previous research works on reconstructing the human face can be categorized based on the approaches used, including vision-based, wearable sensor-based, speech-based, and RF (Radio Frequency) signal-based methods. The vision-based solution leverages the technique in computer vision to

generate the facial landmark or the parameter of the facial model [15, 25, 74, 79]. With wearable sensors, existing works [75] collect bursts of electricity when muscle contracts via an Electromyography (EMG) sensor, then build a neural network for reconstructing facial landmarks and the face model. RF signal-based solution mainly uses mmWave sensor to collect the reflected signal from face [78] and reconstruct facial landmarks for expression classification. Similarly, speech-based solutions generate facial animations from acoustic signals [19, 48]. However, all the above works on the reconstruction cannot extract enough features needed for authentication purposes.

• **Face-related Authentication.** Previous face-related authentication can be divided based on the feature used, including contour-based and movement-based. The contour-based method extracts static face contour for authentication via camera [10, 65], mmWave sensor [81], RFID [80], WiFi [27], acoustic signal [91]. However, these methods are vulnerable to replay attack [81] or need liveness detection [37]. Movement-based methods collect face-related motion, such as lip movement when smiling [32] / speaking [40] and speech facial vibration [60]. However, not like mmFaceID, these methods only capture partial information about facial movements and work at short distances.

• **DoA Schemes in Wireless Sensing.** DoA estimation has been employed to generate RF-map in wireless sensing. Previous research can be categorized into two categories: WiFi-based approaches and mmWave-based approaches. In WiFi-based approaches, a customized complex antenna array is connected to WiFi device to estimate 2D DoA [51, 88], which is costly and impractical. mmWave-based approaches leverage COTS mmWave sensor to estimate DoA [31, 78, 82, 83]. However, they can only achieve 1D DoA due to the limitation of the antenna array, which causes needed information lost.

11 Conclusion

In this paper, we present mmFaceID, a novel mmWave-based user authentication system leveraging dynamic human facial activities. When performing word reading, mmFaceID reconstructs facial activities via estimated expression parameters, in which unique features can be extracted for user authentication. We have theoretically proved the high entropy of using facial activity as biometrics. Comprehensive experiments involving human subjects show that mmFaceID could achieve high accuracy and is robust to distance, orientation, occlusions, and language background.

Acknowledgement

We would appreciate the efforts of all anonymous reviewers and the shepherd who helps improve the quality of this paper. The work of L. Guo is partially supported National Science Foundation under grant CNS-2008049, CCF-2312616, CCF-2427875, and CNS-2431440. The work of X. Ma is partially supported by National Science Foundation under grant CCF-2427875.

Appendix

A ARTIFACT

The research artifacts accompanying this paper are available via <https://doi.org/10.5281/zenodo.13888475>

References

- [1] Andy Adler, Richard Youmaran, and Sergey Loyka. 2009. Towards a measure of biometric feature information. *Pattern Analysis and Applications* 12, 3 (Sept. 2009), 261–270.
- [2] Sharifah Mumtazah Syed Ahmad, Borhanuddin Mohd Ali, and Wan Azizun Wan Adnan. 2012. Technical issues and challenges of biometric applications as access control tools of information security. *international journal of innovative computing, information and control* 8, 11 (2012), 7983–7999.
- [3] Ildus Ahmetov, Olga Vinogradova, and Alun Williams. 2012. Gene Polymorphisms and Fiber-Type Composition of Human Skeletal Muscle. *International journal of sport nutrition and exercise metabolism* 22 (May 2012), 292–303.
- [4] Neamah Al-Naffakh, Nathan Clarke, Fudong Li, and Paul Haskell-Dowland. 2017. Unobtrusive Gait Recognition Using Smartwatches. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 1–5. ISSN: 1617-5468.
- [5] Constantine A. Balanis. 2016. *Antenna theory: analysis and design*. John Wiley & sons.
- [6] Michela Balconi and Giulia Fronda. 2021. How to Induce and Recognize Facial Expression of Emotions by Using Past Emotional Memories: A Multimodal Neuroscientific Algorithm. *Frontiers in Psychology* 12 (2021).
- [7] Boris Bentsianov and Andrew Blitzer. 2004. Facial anatomy. *Clinics in Dermatology* 22, 1 (Jan. 2004), 3–13.
- [8] Francesco Bergadano, Daniele Gunetti, and Claudia Picardi. 2002. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security* 5, 4 (Nov. 2002), 367–397.
- [9] Kevin W. Bowyer and James S. Doyle. 2014. Cosmetic contact lenses and iris recognition spoofing. *Computer* 47, 5 (2014), 96–98. Publisher: IEEE.
- [10] Hakan Cevikalp and Golar Ghorban Dordinejad. 2019. Discriminatively Learned Convex Models for Set Based Face Recognition. 10123–10132.
- [11] Muralidhar Reddy Challa, Abhinav Kumar, and Linga Reddy Cenkaramaddi. 2021. Face Recognition using mmWave RADAR imaging. In *2021 IEEE International Symposium on Smart Electronic Systems (ISES)*. 319–322.
- [12] 3D People Counting. [n. d.]. https://dev.ti.com/tirex/explore/node?a=1AslXXD_1.00.00.26&node=A_AKE9yB4dmaGBeIUYsUEHog_radar_toolbox_1AslXXD_1.00.00.26
- [13] G. De Bonneau, S. Vergez, B. Chaput, B. Vairel, E. Serrano, E. Chantalat, and P. Chaynes. 2019. Variability in facial-muscle innervation: A comparative study based on electrostimulation and anatomical dissection. *Clinical Anatomy (New York, N.Y.)* 32, 2 (March 2019), 169–175.
- [14] Lang Deng, Jianfei Yang, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2022. GaitFi: Robust Device-Free Human Identification via WiFi and Vision Multimodal Learning. arXiv:2208.14326 [cs].
- [15] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Style Aggregated Network for Facial Landmark Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, USA, 379–388.
- [16] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics* 35, 4 (July 2016), 1–11.
- [17] Paul Ekman. 2006. *Darwin and facial expression: A century of research in review*. Ishk.
- [18] Nesli Erdogmus and Sebastien Marcel. 2014. Spoofing face recognition with 3D masks. *IEEE transactions on information forensics and security* 9, 7 (2014), 1084–1097. Publisher: IEEE.
- [19] Sefik Emre Eskimez, Ross K. Maddox, Chenliang Xu, and Zhiyao Duan. 2020. Noise-Resilient Training Method for Face Landmark Generation From Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 27–38. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.
- [20] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom '21)*. Association for Computing Machinery, New York, NY, USA, 337–349.
- [21] Roberto Font, Juan M. Espin, and María José Cano. 2017. Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 Challenge. In *Interspeech*. 7–11.
- [22] Jonah Gamba. 2020. *Radar Signal Processing for Autonomous Driving*. Springer Singapore, Singapore.
- [23] Luca Ghiani, David Yambay, Valerio Mura, Simona Tocco, Gian Luca Marcialis, Fabio Roli, and Stephanie Schuckers. 2013. Livdet 2013 fingerprint liveness detection competition 2013. In *2013 international conference on biometrics (ICB)*. IEEE, 1–6.
- [24] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z. Li. 2021. Towards Fast, Accurate and Stable 3D Dense Face Alignment. arXiv:2009.09960 [cs].
- [25] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z. Li. 2021. Towards Fast, Accurate and Stable 3D Dense Face Alignment. arXiv:2009.09960 [cs].
- [26] Mahyar Hamed, Sh-Hussain Salleh, Chee-Ming Ting, Mehdi Astaraki, and Alias Mohd Noor. 2016. Robust facial expression recognition for MuCI: a comprehensive neuromuscular signal analysis. *IEEE Transactions on Affective Computing* 9, 1 (2016), 102–115. Publisher: IEEE.
- [27] Eran Hof, Amichai Sanderovich, Mohammad Salama, and Evyatar Hemo. 2020. Face Verification Using mmWave Radar Sensor. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIC)*. 320–324.
- [28] Ben Hutchins, Anudeep Reddy, Wenqiang Jin, Michael Zhou, Ming Li, and Lei Yang. 2018. *Beat-PIN: A User Authentication Mechanism for Wearable Devices Through Secret Beats*. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. ACM, Incheon Republic of Korea, 101–115.
- [29] K. Inthavasis and D. Lopresti. 2012. Secure speech biometric templates for user authentication. *IET Biometrics* 1, 1 (2012), 46.
- [30] Kalvik. 2020-07-07T14:52:00Z. Gait Based User Recognition from mmWave Radar Data. original-date: 2020-07-07T14:52:00Z.
- [31] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. m3Track: <u>mm</u>-<u>wave</u>-based <u>m</u>-<u>ulti</u>-user 3D posture tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22)*. Association for Computing Machinery, New York, NY, USA, 491–503.
- [32] Ling Kuang, Fanzi Zeng, Daibo Liu, Hangcheng Cao, Hongbo Jiang, and Jiangchuan Liu. 2023. LipAuth: Securing Smartphone User Authentication with Lip Motion Patterns. *IEEE Internet of Things Journal* (2023). Publisher: IEEE.
- [33] Hao Li, Ruofeng Liu, Shuai Wang, Wencho Jiang, and Chris Xiaoxuan Lu. 2022. Pedestrian Liveness Detection Based on mmWave Radar and Camera Fusion. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. 262–270. ISSN: 2155-5494.
- [34] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, and Kui Ren. 2020. VocalPrint: exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 312–325.
- [35] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. 2021. End-to-End Model-Based Gait Recognition. In *Computer Vision – ACCV 2020*, Hiroshi Ishikawa, Cheng-Lin Liu, Tomas Pajdla, and Jianbo Shi (Eds.). Vol. 12624. Springer International Publishing, Cham, 3–20. Series Title: Lecture Notes in Computer Science.
- [36] Xinrong Li, Xiaodong Wang, Qing Yang, and Song Fu. 2021. Signal Processing for TDM MIMO FMCW Millimeter-Wave Radar Sensors. *IEEE Access* 9 (2021), 167959–167971. Conference Name: IEEE Access.
- [37] Yan Li, Yingjia Li, Qiang Yan, Hancong Kong, and Robert H. Deng. 2015. Seeing Your Face Is Not Enough: An Inertial Sensor-Based Liveness Detection for Face Authentication. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, Denver Colorado USA, 1558–1569.
- [38] Zhengxiong Li, Fenglong Ma, Aditya Singh Rathore, Zhuolin Yang, Baicheng Chen, Lu Su, and Wenya Xu. 2020. WaveSpy: Remote and Through-wall Screen Attack via mmWave Sensing. In *2020 IEEE Symposium on Security and Privacy (SP)*. 217–232. ISSN: 2375-1207.
- [39] Tiantian Liu, Feng Lin, Chao Wang, Chenhan Xu, Xiaoyu Zhang, Zhengxiong Li, Wenya Xu, Ming-Chun Huang, and Kui Ren. 2023. WavoID: Robust and Secure Multi-modal User Identification via mmWave-voice Mechanism. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–15.
- [40] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. LipPass: Lip Reading-based User Authentication on Smartphones Leveraging Acoustic Signals. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. 1466–1474.
- [41] Tania Marur, Yakup Tuna, and Selman Demirci. 2014. Facial anatomy. *Clinics in Dermatology* 32, 1 (Jan. 2014), 14–23.
- [42] Janelle Mason, Rushit Dave, Prosenjit Chatterjee, Ieschechia Graham-Allen, Albert Esterline, and Kaushik Roy. 2020. An investigation of biometric authentication in the healthcare environment. *Array* 8 (2020), 100042. Publisher: Elsevier.
- [43] A. N. Meltzoff and M. K. Moore. 1977. Imitation of facial and manual gestures by human neonates. *Science (New York, N.Y.)* 198, 4312 (Oct. 1977), 74–78.
- [44] David Menotti, Giovanni Chiachia, Allan Pinto, William Robson Schwartz, Helio Pedrini, Alexandre Xavier Falcao, and Anderson Rocha. 2015. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security* 10, 4 (2015), 864–879. Publisher: IEEE.
- [45] Srinivasan Murali, Wenqiang Jin, Vighnesh Sivaraman, Huadi Zhu, Tianxi Ji, Pan Li, and Ming Li. 2023. Continuous Authentication Using Human-Induced Electric Potential. In *Proceedings of the 39th Annual Computer Security Applications Conference (ACSAC '23)*. Association for Computing Machinery, New York, NY, USA, 409–423.
- [46] Rahul Murmuria, Angelos Stavrou, Daniel Barbará, and Dan Fleck. 2015. Continuous Authentication on Mobile Devices Using Power Consumption, Touch Gestures and Physical Movement of Users. In *Research in Attacks, Intrusions, and Defenses (Lecture Notes in Computer Science)*, Herbert Bos, Fabian Monrose, and Gregory Blanc (Eds.). Springer International Publishing, Cham, 405–424.

- [47] Obi Ogbanufe and Dan J. Kim. 2018. Comparing fingerprint-based biometrics authentication versus traditional authentication methods for e-payment. *Decision Support Systems* 106 (2018), 1–14. Publisher: Elsevier.
- [48] Hai X. Pham, Yuting Wang, and Vladimir Pavlovic. 2017. End-to-end Learning for 3D Facial Animation from Raw Waveforms of Speech. arXiv:1710.00920 [cs].
- [49] Nataasha Raul, Radha Shankarmani, and Padmaja Joshi. 2020. A Comprehensive Review of Keystroke Dynamics-Based Authentication Mechanism. In *International Conference on Innovative Computing and Communications*, Ashish Khanna, Deepak Gupta, Siddhartha Bhattacharyya, Vaclav Snasel, Jan Platos, and Aboul Ella Hassanien (Eds.). Vol. 1059. Springer Singapore, Singapore, 149–162. Series Title: Advances in Intelligent Systems and Computing.
- [50] Yanzhi Ren, Siyi Li, Chen Chen, Hongbo Liu, Jiadi Yu, Yingying Chen, Haomiao Yang, and Hongwei Li. 2023. Robust Indoor Location Identification for Smartphones Using Echoes From Dominant Reflectors. *IEEE Transactions on Mobile Computing* (2023), 1–17.
- [51] Yili Ren and Jie Yang. 2022. Robust Person Identification: A WiFi Vision-based Approach. arXiv:2210.00127 [cs].
- [52] Napa Sae-Bae and Nasir Memon. 2022. Distinguishability of keystroke dynamic template. *Plos one* 17, 1 (2022), e0261291. Publisher: Public Library of Science San Francisco, CA USA.
- [53] Nikolaus P. Schumann, Kevin Bongers, Hans C. Scholle, and Orlando Guntinas-Lichius. 2021. Atlas of voluntary facial muscle activation: Visualization of surface electromyographic activities of facial muscles during mimic exercises. *PLOS ONE* 16, 7 (July 2021), e0254932. Publisher: Public Library of Science.
- [54] Aqara Presence Sensor. [n. d.]. . <https://www.aqara.com/us/product/presence-sensor-fp2/>
- [55] Joseph Shelton, Kaushik Roy, Brian O'Connor, and Gerry V. Dozier. 2014. Mitigating iris-based replay attacks. *International Journal of Machine Learning and Computing* 4, 3 (2014), 204. Publisher: IACSIT Press.
- [56] Xingjian SHI, Zhoulong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc.
- [57] J. A. Simoneau and C. Bouchard. 1995. Genetic determinism of fiber type proportion in human skeletal muscle. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 9, 11 (Aug. 1995), 1091–1095.
- [58] Jasvinder Pal Singh, Sanjeev Jain, Sakshi Arora, and Uday Pratap Singh. 2018. Vision-Based Gait Recognition: A Survey. *IEEE Access* 6 (2018), 70497–70527.
- [59] Daniel F. Smith, Arnold Willem, and Brian C. Lovell. 2015. Face recognition on consumer devices: Reflections on replay attacks. *IEEE Transactions on Information Forensics and Security* 10, 4 (2015), 736–745. Publisher: IEEE.
- [60] Tanmay Srivastava, Phuc Nguyen, Shijia Pan, and Shubham Jain. 2023. Jawthenticate: Microphone-free Speech-based Authentication using Jaw Motion and Facial Vibrations. *free Speech* (2023).
- [61] Ian Stavness, Mohammad Ali Nazari, Pascal Perrier, Didier Demolin, and Yohan Payan. 2013. A biomechanical modeling study of the effects of the orbicularis oris muscle and jaw posture on lip shape. *Journal of speech, language, and hearing research: JSLHR* 56, 3 (June 2013), 878–890.
- [62] Bharath Sudharsan, Peter Corcoran, and Muhammad Intizar Ali. 2022. Smart speaker design and implementation with biometric authentication and advanced voice interaction capability. arXiv:2207.10811 [cs, eess].
- [63] Yagiz Sutcu, Elham Tabassi, Husrev T. Sencar, and Nasir Memon. 2013. What is biometric information and how to measure it?. In *2013 IEEE International Conference on Technologies for Homeland Security (HST)*. 67–72.
- [64] Kenta Takahashi and Takao Murakami. 2014. A measure of information gained through biometric systems. *Image and Vision Computing* 32, 12 (2014), 1194–1203. Publisher: Elsevier.
- [65] K. Venkataramani, S. Qidwai, and B.V.K. Vijayakumar. 2005. Face authentication from cell phone camera images with illumination and temporal variations. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35, 3 (Aug. 2005), 411–418. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews).
- [66] Ding Wang, Haibo Cheng, Ping Wang, Xinyi Huang, and Gaopeng Jian. 2017. Zipf's law in passwords. *IEEE Transactions on Information Forensics and Security* 12, 11 (2017), 2776–2791.
- [67] Ding Wang, Qianchen Gu, Xinyi Huang, and Ping Wang. 2017. Understanding Human-Chosen PINs: Characteristics, Distribution and Security. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM, Abu Dhabi United Arab Emirates, 372–385.
- [68] Shuai Wang, Luoyi Mei, Zhimeng Yin, Hao Li, Ruofeng Liu, Wenchao Jiang, and Chris Xiaoxuan Lu. 2023. End-to-End Target Liveness Detection via mmWave Radar and Vision Fusion for Autonomous Vehicles. *ACM Transactions on Sensor Networks* (Oct. 2023), 3628453.
- [69] Xuan Wang and Jiro Tanaka. 2018. GeSID: 3D Gesture Authentication Based on Depth Camera and One-Class Classification. *Sensors* 18, 10 (Oct. 2018), 3265. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.
- [70] Qi Wei. 2023. Association between the PPARGC1A Gly482Ser polymorphism and muscle fitness in Chinese schoolchildren. *PloS One* 18, 4 (2023), e0284827.
- [71] Julie D. White, Karlijne Indencleef, Sahin Naqvi, Ryan J. Eller, Hanne Hoskens, Jasmien Roosenboom, Myoung Keun Lee, Jiarui Li, Jaaved Mohammed, Stephen Richmond, Ellen E. Quillen, Heather L. Norton, Eleanor Feingold, Tomek Swigut, Mary L. Marazita, Hilde Peeters, Greet Hens, John R. Shaffer, Joanna Wysocka, Susan Walsh, Seth M. Weinberg, Mark D. Shriver, and Peter Claes. 2021. Insights into the genetic architecture of the human face. *Nature Genetics* 53, 1 (Jan. 2021), 45–53. Number: 1 Publisher: Nature Publishing Group.
- [72] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. arXiv:1807.06521 [cs].
- [73] Libing Wu, Jingxiao Yang, Man Zhou, Yanjiao Chen, and Qian Wang. 2020. LVID: A Multimodal Biometrics Authentication System on Smartphones. *IEEE Transactions on Information Forensics and Security* 15 (2020), 1572–1585. Conference Name: IEEE Transactions on Information Forensics and Security.
- [74] Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan. 2016. Facial Landmark Detection with Tweaked Convolutional Neural Networks. arXiv:1511.04031 [cs] (March 2016). arXiv: 1511.04031.
- [75] Yi Wu, Vimal Kakaraparthi, Zhuohang Li, Tien Pham, Jian Liu, and Phuc Nguyen. 2021. BioFace-3D: continuous 3d facial reconstruction through lightweight single-ear biosensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. ACM, New Orleans Louisiana, 350–363.
- [76] Zhizheng Wu, Sheng Gao, Eng Siong Cling, and Haizhou Li. 2014. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2014 Asia-Pacific. IEEE, 1–5.
- [77] Naiqi G. Xiao, Steve Perrotta, Paul C. Quinn, Zhe Wang, Yu-Hao P. Sun, and Kang Lee. 2014. On the facilitative effects of face motion on face recognition and its development. *Frontiers in Psychology* 5 (June 2014).
- [78] Jiahong Xie, Hao Kong, Jiadi Yu, Yingying Chen, Linghe Kong, Yanmin Zhu, and Feilong Tang. 2023. mm3DFace: Nonintrusive 3D Facial Reconstruction Leveraging mmWave Signals. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. ACM, Helsinki Finland, 462–474.
- [79] Xuehan Xiong and Fernando De la Torre. 2013. Supervised Descent Method and Its Applications to Face Alignment. 532–539.
- [80] Weiye Xu, Jianwei Liu, Shimin Zhang, Yuanqing Zheng, Feng Lin, Jinsong Han, Fu Xiao, and Kui Ren. 2021. RFace: Anti-Spoofing Facial Authentication Using COTS RFID. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*. 1–10. ISSN: 2641-9874.
- [81] Weiye Xu, Wenfan Song, Jianwei Liu, Yajie Liu, Xin Cui, Yuanqing Zheng, Jinsong Han, Xinhui Wang, and Kui Ren. 2022. Mask does not matter: anti-spoofing face authentication using mmWave without on-site registration. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking (MobiCom '22)*. Association for Computing Machinery, New York, NY, USA, 310–323.
- [82] Hongfei Xue, Qiming Cao, Yan Ju, Haochen Hu, Haoyu Wang, Aidong Zhang, and Lu Su. 2023. M4esh: mmWave-Based 3D Human Mesh Construction for Multiple Subjects. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22)*. Association for Computing Machinery, New York, NY, USA, 391–406.
- [83] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21)*. Association for Computing Machinery, New York, NY, USA, 269–282.
- [84] Xin Yang, Song Yang, Jian Liu, Chen Wang, Yingying Chen, and Nitesh Saxena. 2021. Enabling finger-touch-based mobile user authentication via physical vibrations on IoT devices. *IEEE Transactions on Mobile Computing* 21, 10 (2021), 3565–3580. Publisher: IEEE.
- [85] R. Youmaran and A. Adler. 2012. Measuring biometric sample quality in terms of biometric feature information in iris images. *Journal of Electrical and Computer Engineering* 2012 (Jan. 2012), 22:22.
- [86] Xiaojing Yu, Zhijun Zhou, Mingxue Xu, Xuanke You, and Xiang-Yang Li. 2020. ThumbUp: Identification and Authentication by Smartwatch using Simple Hand Gestures. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 1–10. ISSN: 2474-249X.
- [87] Yaoyuan Yuan, Jizhong Zhao, Wei Xi, Chen Qian, Xiaobin Zhang, and Zhi Wang. 2017. SALM: Smartphone-Based Identity Authentication Using Lip Motion Characteristics. In *2017 IEEE International Conference on Smart Computing (SMART-COMP)*. 1–8.
- [88] Feng Zhang, Chenshu Wu, Beibei Wang, and K. J. Ray Liu. 2021. mmEye: Super-Resolution Millimeter Wave Imaging. *IEEE Internet of Things Journal* 8, 8 (April 2021), 6995–7008. Conference Name: IEEE Internet of Things Journal.
- [89] Yanchao Zhao, Ran Gao, Shangqing Liu, Lei Xie, Jie Wu, Huawei Tu, and Bing Chen. 2021. Device-Free Secure Interaction With Hand Gestures in WiFi-Enabled IoT Environment. *IEEE Internet of Things Journal* 8, 7 (April 2021), 5619–5631.
- [90] Kai Zheng, Wuqiong Zhao, Timothy Woodford, Renjie Zhao, Xinyu Zhang, and Yingbo Hua. 2024. Enhancing mmWave Radar Sensing Using a Phased-MIMO Architecture. In *Proceedings of the 22nd Annual International Conference on Mobile*

- Systems, Applications and Services (MOBISYS '24)*. Association for Computing Machinery, New York, NY, USA, 56–69.
- [91] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2018. EchoPrint: Two-factor Authentication using Acoustics and Vision on Smartphones. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18)*. Association for Computing Machinery, New York, NY, USA, 321–336.
 - [92] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. Visemenet: audio-driven animator-centric speech animation. *ACM Transactions on Graphics* 37, 4 (Aug. 2018), 1–10.
 - [93] Huadi Zhu, Wenqiang Jin, Mingyan Xiao, Srinivasan Murali, and Ming Li. 2020. BlinkKey: A Two-Factor User Authentication Method for Virtual Reality Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 1–29.
 - [94] Huadi Zhu, Mingyan Xiao, Demoria Sherman, and Ming Li. 2023. SoundLock: A Novel User Authentication Scheme for VR Devices Using Auditory-Pupillary Response. In *Proceedings 2023 Network and Distributed System Security Symposium*. Internet Society, San Diego, CA, USA.
 - [95] Maheen Zulfiqar, Fatima Syed, Muhammad Jaleed Khan, and Khurram Khurshid. 2019. Deep face recognition for biometric authentication. In *2019 international conference on electrical, communication, and computer engineering (ICECCE)*. IEEE, 1–6.